DOCUMENT RESUME

ED 174 649 TM 009 498

AUTHOR Warren, Jonathan R.
TITLE The Measurement of Academic Competence.
INSTITUTION Educational Testing Service, Berkeley, Calif.
SPONS AGENCY Fund for the Improvement of Postsecondary Education (DHEW), Washington, D.C.
PUB DATE Dec 78
GRANT G007603526
NOTE 83p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Academic Achievement; Academic Standards; *Basic Skills; College Teachers; Communication Skills; *General Education; Higher Education; Interpersonal Competence; Performance Based Education; *Response Style (Tests); Scoring; *Teacher Participation; *Test Construction; Test Reliability; Test Validity; *Thought Processes; Timed Tests

ABSTRACT

Faculty from fifteen colleges developed open-ended questions to assess group competence in four areas of learning associated with general education--communication skill, analytic thinking, synthesizing ability, and social/cultural awareness. A secondary purpose was to determine faculty expectations for learning in the general education component of bachelor's degrees and to compare the effects of different forms of general education curricula. The questions were scored by the faculty by sorting responses into four to seven categories, each indicating a different way students may not reach a completely successful response. Results are reported for the final forty-seven questions administered and for a subset of thirty-two considered most valid as a measure of student performance. The questions can be scored with acceptable levels of interrater reliability and answered within the time requirement of ten minutes per question. To combine results from several questions into inferences about performance with respect to a common competence, however, requires evidence that the questions do in fact measure the same competence. Since evidence of that was not present, comparisons of faculty expectations and curricular forms were not possible. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174649

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Measurement of Academic Competence

Jonathan R. Warren
Senior Research Scientist
Educational Testing Service
1947 Center Street
Berkeley, California 94704
(415) 849-0950

December 1978

Sponsored by the
Fund for the Improvement of Postsecondary Education
under
Grant Number G007603526

Project starting date: July 1, 1976
Project ending date: December 31, 1977
Duration: 18 months

TM009 498

Executive Summary

The project goals were to develop free-response questions to assess competence in four areas of learning associated with general education--communication skill, analytic thinking, synthesizing ability, and social/cultural awareness--and to use the developed questions to determine faculty expectations for learning in the general education component of bachelor's degrees and to compare the effects of different forms of general education curricula. Fifteen four-year colleges and universities in California--large and small, public and private, selective and nonselective--their faculties, and a total of about 3,000 students participated. Faculty members wrote questions that students could answer in less than 10 minutes each, that were expected to reveal different levels of performance in the four general education competences defined for the project, and that could be scored by sorting the responses into four to seven categories. Two trial administrations of the questions at the 15 institutions were followed by a third administration to students classified according to the type of general education curriculum followed.

The type of question developed attracted the interest of faculty and students alike. The questions seemed to both groups to be realistic indicators of the kinds of broad intellectual skills colleges are expected to produce. They can be scored expeditiously with acceptable levels of reliability, and the time requirements are feasible for use of the questions to assess the performance levels of groups of students--that is, to assess program or curricular effects on student learning rather than individual student performance. To combine results from several questions into

inferences about performance with respect to a common competence, however, requires evidence that the several questions do in fact measure the same competence. That evidence in the present project was not sufficiently strong or extensive to permit such inferences. As a consequence, comparisons of the effects of different curricular forms and determination of faculty expectations were not possible.

One major result of the project is the identification of several components within each of the four competences assessed, each component indicating a different way students may not reach a completely successful response. The observation of different kinds of inadequacies among the students in a program should permit modification of the content or instructional procedures to improve them.

A second result was the demonstration of a feasible way to assess program effectiveness with respect to subtle, complex kinds of learning. Although faculty at several institutions outside the project where the questions were tried reported that the scoring process for some questions was too complex and that some questions were too difficult for poorly prepared students, the questions in general and their use in program evaluation were viewed favorably.

The scoring of the questions by faculty members on a campus and the associated observations of types of student performance helped faculty members clarify their own definitions of the kinds of learning to which general education should be addressed and showed which deficiencies were most prevalent in their students. Formulating new or revised questions of a similar kind directed to predetermined goals was also an activity

faculty members found stimulating and informative.

Finally, a group of questions has been developed, each of which provides for broad discrimination among students. While their interpretation requires further confirmation, the differences in response that they elicit seem to be clearly pertinent to four broad goals of general education.

The Measurement of Academic Competence

For the past hundred years, ever since colleges and universities began offering students a choice among fields of study, a conflict has existed in higher education over the relative emphasis to be given specialized study in a selected major field as opposed to broad study in the liberal arts and sciences. Both kinds of study are almost invariably required for bachelor's degrees in the United States. Yet while curricular requirements within particular fields are well-established, those designed for broad, general studies are not.

The appropriate form and content of the general part of the bachelor's degree program are periodically re-examined. A major effort to define the general curriculum has recently been completed at Harvard University about 30 years after an earlier Harvard committee published the widely discussed Harvard "Red Book" on approximately the same issue. Soon after the earlier Harvard report, the American Council on Education completed a major study to define the goals of general education and develop objective measures of the extent to which they are achieved. Neither the definitions nor the measure had much impact, perhaps because the launching of Sputnik by the Soviet Union focused the attention of those in higher education on the specifics of science and engineering programs at the expense of the general goals of higher education.

Daniel Bell (1966) in the 1960's and Sidney Hook (1975) in the 1970's have represented Columbia University in attempts to define the purposes and appropriate curricular structures for undergraduate programs. Both Bell and Hook seem to view higher education from the perspective of the seletive liberal arts college, preparing an intellectually elite student body for schol-

arly work in graduate school and beyond. The Harvard Red Book had the same bias but with slightly more justification at a time when the entry into college of as many as half of the high school graduates seemed a barely realistic expectation.

A more serious limitation of every examination of the general education curriculum, with the single exception of the ACE study, has been the failure to relate desired goals to instructional or pedagogical realities. Adjusting the content of curricula or course syllabi, which is the action toward which the various studies point, may change the books students read and the topics around which lectures and seminars are organized. Without more explicit attention to the activities rather than the topics of the classroom and to the assesment of desired kinds of student learning rather than accumulation of detailed knowledge, curricular adjustments are not likely to produce learning different in any lasting way from the learning that occurs in any of the traditionally organized programs.

Students learn the specifics of the courses they take. When a number of courses in a common field are interrelated with respect to content, students develop an understanding in some depth of the content and procedures of a given field. But the goals of general education are less explicit than those of a subject-matter field, and their accomplishment tends to be assumed if only the subject matter studied is diverse enough. Bringing course content into line with the most recent applications or theoretical positions in the field, which is the usual consequence of curricular studies, does not necessarily change the general nature of student learning, which is the implicit if not explicit goal of studies of the curriculum in general.

The utility of measures of general learning

If what many perceive as too extreme a swing in recent years toward specialized learning at the expense of general learning is to be countered, the accomplishments of general education will need to be demonstrated. Vague statements about the need for breadth, or that students given the tools for learning will be equipped for further learning as it becomes needed, or other platitudes of liberal arts colleges and general education programs will not be sufficient to resist two commonly heard arguments. The first is that the explosion of knowledge requires increasingly concentrated study if a student is to become competent in a field. The second is that the economic benefits of technical competence are too great to pass up for a little dabbling in other fields. The obvious accomplishment, to the student and to others, of relatively advanced knowledge in a field has great appeal. So too does the sense of purpose associated with preparation for some kind of occupation, however nebulous the connection may be between a major field of study and a post-graduation job. To counter those powerful pulls toward specialization, general education must demonstrate concrete results. The present project was planned to provide measures with which programs of general education could demonstrate their merit.

The development of measures of general learning should lead to several other useful results. First, it should permit standards in general education to be defined and examined. Institutions faced with claims that new and unfamiliar procedures can educate students at least as well as the usual methods will have a way to challenge those claims. And if the claims are legitimate, that too can be demonstrated. Until some basis is available to define in a general sense the standards of accomplishment on which higher

education bases much of its justification for support, claims of eroding standards and the debasing of the degree, and their counterclaims, will be little more than cocktail talk. Instructional and curricular modifications will appear, will themselves be modified, and little will be known of their effects beyond the satisfaction of the sense of venture that produced them.

Problems in measuring general learning

The measurement of intellectual performance is a peculiarly disjointed field. Almost all the theoretical work on intellectual performance and its measurement is concerned with comparatively stable aspects of performance, such as reasoning, verbal fluency, comprehension, and numerical facility, which are resistant to short-term educational change.

Research on instructional methods typically gives attention to the conditions of instruction and to stable characteristics of the learners. The material to be learned is often extremely circumscribed, at times highly contrived, and rarely related to anything faculty members are likely to describe as general goals of their courses. Finally, most of the commercial investment in the measurement of intellectual performance is directed to achievement in specific areas of study which, particularly in higher education, reflect the content of a course of study rather than the more general kinds of intellectual performance which many faculty members value. These include such activities as identifying the critical issues in an argument, discriminating between strongly supported and purely conjectural assertions, and forming generalizations from a number of instances. A similar observation has been made with respect to learning from written prose. "The learning and retention of facts have been studied, but not many studies ask how these facts are organized by the reader. Making inferences and drawing conclusions from written passages has been little studied

(Glaser and Resnick, 1972)." Thus basic intellectual abilities resistant to educational efforts are studied; instructional procedures are studied in their effects on comparatively trivial kinds of learning in the undemonstrated expectation that the results will apply to more important kinds of learning; academic achievement is measured largely in terms of factual content rather than in the exercise of the intellectual skills acquired in an educational program.

A major problem in prior attempts to develop measures of general learning has been the relationship between measures of comparatively stable intellectual abilities, or aptitude, and acquired abilities, or achievement. Achievement is necessarily dependent both on the intellectual resources, or aptitude, students bring to bear on learning and on the nature of the particular course or experience in which the learning occurs. Separating these two influences on achievement has been difficult. Most achievement measures that are sufficiently independent of aptitude to reflect recent learning are too closely tied to the specific content of the instruction to assess general forms of intellectual accomplishment. Measures that have been somewhat freer of specific content have been difficult to distinguish from measures of aptitude.

The long history of failure to develop measures of academic achievement that are broader than course content, the kind of achievement often described as the objectives of general education, makes the success of new efforts problematical. Yet without such measures, the goals of college curricula are likely to continue to narrow, a process that Blackburn et al. (1976) have documented. Broad academic objectives that cannot be observed in student learning will lose ground to the specific objectives

of course syllabi that can be observed--learning to approximate the solution to an equation through a Fourier expansion or to describe Virginia Woolf's use of irony in Orlando.

Purposes and Procedures

Despite the difficulties in developing measures of academic abilities or competences that are acquired or developed more fully in college but that are also generalizable beyond the content or experiences of particular courses, that was the primary purpose of the project. Secondary purposes were to collect faculty judgments of appropriate standards to apply to those competences and to demonstrate their use in comparing different curricular structures.

Two recent developments suggested that a new attempt to devise measures of general academic learning that would be distinct from aptitude measures might have some chance of success. A study that immediately preceded the present one showed little coherence or purpose in baccalaureate programs outside the student's major fields of study. At the same time, while many faculty members were clear in their desire to incorporate broad intellectual objectives into their courses, students reported little or no growth in those areas. The collective perceptions of several hundred faculty members led to the identification of about ten generally accepted objectives of baccalaureate programs that went beyond course content (Warren, 1976). The clarity of definition and wide acceptance of these objectives of general education and the absence of effective measures of their accomplishment suggested that a new attempt to develop appropriate measures would be worth at least one more attempt.

Some preliminary success with the use of open-ended questions to measure broad aspects of learning in graduate psychology programs (Frederiksen & Ward, 1978) was the second development that gave some

impetus to the present project. The multiple-choice tests commonly used to assess both aptitude and achievement have long been criticized as inherently unsuited to the measurement of subtle aspects of learning. An examination task that requires the student to construct a response is believed by many to be a more promising approach to the measurement of complex kinds of intellectual accomplishment. Freely constructed responses, however, are difficult to score reliably and require so much response time that comparatively little information about each student can be provided within reasonable time limits. Frederiksen and Ward (1978) provided one answer to the scoring problem, and limiting the measurement task to program rather than individual assessment reduced the problem of low yields of information per student tested.

Frederiksen and Ward's procedure for making the scoring of open-ended questions reasonably objective and uniform is to define a set of categories to which responses can be assigned on the basis of some fairly clear characteristic of the response. Since the scorer is required only to sort responses by type and need not assign a value to them, the task can be done quickly and, for well developed questions and categories, reliably. In a separate process, knowledgeable persons determine a value for each category, and each response assigned to a given category receives the value of that category.

The comparatively low amount of information that can be gathered about the performance of each student because of the length of time required for each question (about 10 minutes in the present project) is not a problem if the performance of groups rather than individual students is of interest. When the concern is whether a new program produces

better results than the customary one, or whether the desired results are being achieved by either program, group rather than individual performance must be measured. Having information about a moderately large group of students permits the information about each individual student to be limited.

The procedure followed was therefore to develop open-ended questions that could be answered in five to ten minutes and that could be scored by a comparatively untrained person by assigning each response to one of a set of predefined categories. But useful questions don't appear spontaneously, and specifications had to be stated that would guide the writing of the individual questions.

Specification of the nature of the questions

A committee of 16 faculty members, one from each of 16 colleges and universities in California--large and small, public and private, selective and nonselective--met to identify a limited number of competences, clearly pertinent to the purposes of general education, that could be defined with enough precision to guide the writing of questions. The prior study (Warren, 1976) in which a large number of descriptive statements of college student performance (e.g., writes critically) had been organized into about 10 more general kinds of performance (e.g., analytic ability) provided a starting point. The committee selected three of these areas of achievement--communication, analytic thinking, and synthesizing ability--and clarified their definitions. They added a fourth, awareness, that had appeared in a different part of the earlier study in a series of interviews with faculty members about course objectives. These four types of student performance or

competence are defined below, with more detailed definitions attached as Appendix A.

Communication. Clarity and precision of written expression; ready movement from one mode of written expression to another.

Analytic Thinking. Identification of the essential components of ideas, of inconsistencies and inaccuracies in the statement of ideas; discrimination of fact from conjecture, the important from the unimportant, the relevant from the irrelevant.

Synthesizing ability. Generalization of ideas to new or broader contexts; integration of given knowledge or information into more general structures.

Awareness. Sensitivity to the implications of thoughts and actions for the social good, and to the importance of values and cultural contexts in human affairs.

The committee also decided that questions to measure these four kinds of competence should involve content in four academic areas--natural sciences and mathematics, humanities, social sciences, and history and political science. The separation of history and political science from the humanities and social science was prompted by the prominence of U.S. history and government in the general education requirements of most colleges and universities.

The questions were to provide a way of assessing general academic competences that would reflect the broad learning expected of college graduates. Their answers could not require knowledge likely to be gained only in particular courses, yet intellectual skills are not demonstrated without reference to content of some kind. The questions

were therefore designed to draw on the four broad content areas listed above but to assume possession by the students only of knowledge likely to be acquired in a variety of different courses or in the general exchange of ideas expected to occur during most four-year college programs. Most specific information required was therefore given with the question. Answers were scored according to what the students did with the information, how they used it to produce a response of the type requested; they were not asked to produce information.

An appropriate level of difficulty was not easy to achieve. Faculty members are accustomed to writing questions that will discriminate among students who have all studied a particular body of content. Questions were needed that were not trivial, that would challenge most students, including many whose majors were in the content on which a question was based, but that could also be answered by competent students who had not studied in the question's field of content. The questions should also be difficult enough to discriminate among college seniors but not so difficult that freshmen and sophomores would find them totally beyond their reach.

The questions were intended to present students with a well-defined intellectual task for which they had to construct an appropriate response. To allow sufficient time for students to read and understand what is being asked, organize their thoughts, and write out a response, at least five and frequently ten minutes seemed desirable. To get enough information in a reasonable period of time to make the process feasible, no more than ten minutes per question seemed permissible. The questions were therefore designed so the

necessary information and the question could be presented on the upper half of a sheet of paper, with the lower half available for the student's written response.

Trial questions

The 16 faculty members who constituted the committee advising the project plus additional faculty members they recruited on their own campuses wrote a total of approximately 100 questions to meet the specifications described above. From these, four questions were selected from each of the four content areas to measure each of the four competences, a total of 64 questions. Those rejected seemed not clearly related to one of the four competences, vague with respect to the kind of response desired of the student, too difficult, or too easy. The 64 selected questions were administered in groups of eight to about 900 students at 15 participating colleges and universities, one having withdrawn because of the time requirements of the project.

An immediate conclusion as a result of the first trial was that eight questions were too many to be completed comfortably in a fifty-minute class hour. Six seemed more reasonable and turned out in later administrations to be about right.

Students' comments, which were asked for on the cover sheet of the sets of questions, were mixed. Many students were pleased with the novelty of the questions, their challenge, and their scope. They expressed satisfaction at being forced to think in ways quite different from those characteristic of classroom or multiple-choice tests. Other students expressed dismay at the

difficulty of the questions and their inability to deal with them in a way that seemed adequate. The dominant reaction of the students, however, was one of a sense of challenge and satisfaction.

Scoring

The same group of faculty members who had written the questions, with a few more as well as ETS staff, devised categories for scoring the questions. This was done by reading 20 responses to a question and sorting them into four to eight groups that seemed to accommodate the distinctions observed in the responses that were pertinent to the competence being assessed. Poor diction, for example, was used to distinguish between categories of response for questions measuring communication but not between categories in any of the other three competences. Nor did grammatical correctness or prose style enter judgments about responses in competences other than communication. Similarly, the identification of the key elements in a question, indicative of analytic skill, was as far as possible excluded from the formulation of response categories for questions measuring synthesizing ability. This kind of distinction was often difficult to make, and the most obvious ways of discriminating among the responses to a question were at times not clearly associated with the competence intended to be measured.

After one person, usually not the person who had written the question, had developed a set of categories into which 20 responses could sensibly be sorted, two other people tried those categories on a new set of 20 questions. They checked with each other after

all 20 had been scored, looked for the reasons for their disagreements, and revised the categories to reduce those sources of disagreement. Those revised categories were then reviewed by the person who had written the question and revised again if some important distinction in the responses had been lost. All the responses to a question were then assigned to one of the revised categories.

Problems faced in trying to assign responses to one of four to eight categories often were the result of lack of clarity in the wording of the question. Poor discriminability among responses also suggested ways questions could be improved. Students at times suggested ways the questions might be improved. All these sources of information about the clarity of the questions, their pertinence to the competence being measured, and their ability to produce discriminable responses were drawn on in revising the questions and writing new ones for the second trial.

In devising categories into which responses to a question could be usefully sorted, the faculty members were asked to be concerned only with qualitative differences in the responses. Whether one response was better or worse than another was of no immediate concern unless that judgment of relative merit suggested a categorical distinction that could be applied to other questions. The categories were to indicate different types of response without regard for relative merit. In actuality, judgments of merit were difficult to avoid, and almost invariably the person who devised the final set of categories ordered them with respect to his or her view of their relative merit. At times the scorers were disturbed at being forced to

place two responses clearly different in merit into the same category. That difference in the two responses then suggested a revision of the category definitions or perhaps a new category. All the responses to that question that had already been scored then had to be rescored.

Revised questions

The 64 questions administered during the first trial were examined to select a smaller group for more intensive study. Twenty questions were selected to provide as nearly as possible the following characteristics:

1. broad distribution of responses across the scoring categories,
2. moderately reliable scoring,
3. clear representation of each of the four competences.

The purpose of the second trial was to clarify the relationships among a small number of questions that could then act as markers for interpreting the remaining questions. Each of about 500 students in the same 15 institutions completed 6 questions. A final group of 48 questions, some of them revised a second time and with one or two new ones added, were administered to about 1,200 students selected from the 15 participating institutions to represent different types of general education curricula.

One of these final questions produced responses that could not be categorized reliably, and it was not scored. Results are therefore reported for the final 47 questions administered and scored or for a subset of 32 questions that were considered the ones most promising as measures of student performance. These 32 are described in Appendix B

in terms of the task set for the student and the major categories into which the responses are sorted.

Characteristics of the Questions

Reliability

Open-ended questions present several problems, such as assuring that the responses are pertinent or permit inferences of the kind desired, producing a sufficient number of responses in a limited period of time, and scoring the responses reliably. The pertinence of the inferences, or validity, is discussed in a later section. Scoring by assigning a response to one of a set of well-defined categories reduces the scoring time required to a manageable level. Whether it also provides an adequate degree of reliability in the scoring depends on how the information in the assigned score is to be used. The reliability of individual students' responses is a separate issue.

In the usual testing programs scores are assigned to individual students, and moderately large numbers of test items or questions are needed for each test if scores are to be sufficiently reliable for judgments to be made about the performance of individual students. For questions that require about 10 minutes each for a response, several hours of testing time would probably be needed to get one reliable score for an individual student, assuming the usual degree of reliability for individual questions. If, however, a measure of the collective performance of a group of students is desired, the testing time can be reduced as the number of students in the group is increased. Fifty students, for example, each responding to one question, would provide a reliable estimate of the average performance of the group if the questions were interrelated

and had very modest individual reliability. In an hour of testing time, 50 students randomly selected from a larger group could provide reliable estimates of the collective ability of the larger group with respect to as many as six different types of performance. When program effectiveness rather than individual performance is of interest, open-ended questions that take as long as 10 minutes per question for a response and have combined response and scorer reliabilities on the order of .20 can provide quite reliable measures with moderate testing times if the different questions produce comparable types of responses.

The reliability of the scores assigned to open-ended questions can be broken into two components that account for the major sources of error in the scores. One is the error in the students' responses, or the variation that would appear if a large number of students of comparable ability responded to the same question. The other is the scorer's error, or the variation that would appear if a large number of scorers read and scored the same response. When group rather than individual scores are of interest, the usual assumption that response errors are independent of ability and normally distributed about a mean of zero implies that errors in the students' responses will not affect the mean score of the group but may affect the variance of the scores. Floor and ceiling effects introduced when scores are provided for each question in terms of a small number of categories or scores per question make that assumption questionable. In those circumstances, measurement errors will be likely to reduce the variance of the scores and shift the observed mean from its true value toward the mid-point of the scale whenever a moderate number of true scores occur near either extreme. If the students' responses

were totally unreliable, any observed differences in group means would be attributable entirely to error.

Because primary interest was in the distributions of student responses to individual questions rather than in multiple-item scores, data from which estimates of response errors could be determined were not collected. The reliability measures refer only to the reliability with which scorers assign responses to categories. For each question, they were based on a common set of 20 responses scored independently by four persons and were calculated in several ways, each with some deficiency.

Intraclass reliability coefficients

A commonly used procedure for estimating the reliability of measures, particularly when the measures of interest are judges' ratings, is the intraclass correlation coefficient (Hays, 1963; Winer, 1972). It is the ratio of the true variance to the total variance of the measure, or of the true variance to the sum of the true variance and error variance. Disagreement is common, however, over the most useful estimates of true and error variance to be used, and their definition depends in part on the kinds of uses desired of the measures.

Intraclass reliabilities, since they are based on variance estimates, assume that the scale of measurement is an equal-interval scale. In the present study, the scoring categories to which scorers assigned responses were ordered with respect to presumed merit and assigned consecutive whole-digit numbers. Equality of the intervals between categories would be wholly fortuitous and quite unlikely.

To place the response categories for each question on a defensible scale, faculty members ranked selected responses to each of 20 questions with respect to merit. Scale values for each response were calculated using Thurstone's assumption that the discrimination among judged values or merit would be greater for responses farther apart on the underlying scale of merit (Torgerson, 1957). Each category was assigned the mean value of its responses, and all responses in a category were assigned that value. If the errors in assigning responses to categories were greater when the categories did not differ much in value, the reliabilities of scores scaled to the judged value of each category rather than to its numerical order would be greater than reliabilities calculated from the numerical order of the categories. Such was not the case. The reliabilities of the scaled scores of questions were lower than those of the nominal scores as often as they were higher, and the decline in reliability using scaled scores was often appreciable. Scaling the categories often reduced true variance as well as error variance, at times producing a reduced reliability coefficient as well.

The reliability coefficients for both scaled and unscaled scores are shown in Table 1 for 32 of the final questions that were considered most promising. Questions for which no scaled score reliability appears were not among the 20 that had been scaled.

The reliability coefficients for unscaled scores range from .14 to .90 with a median value of .52. Coefficients calculated for 50 sets of random data generated to approximate the observed

distributions of scores ranged from -.16 to .15 with a median value of zero. All but two of the coefficients calculated for real data were greater than .20. The assignment of responses to categories is clearly done with accuracy far greater than chance, but the interpretation to be placed on the resulting reliability coefficients is elusive. The proportion of the variance in scores attributable to the inaccuracies of the average judge assigning scores is understandable up to a point but requires confidence in the quality of the scale intervals. Having two judges score each question and then consult with each other to reconcile any differences will reduce the judges' errors appreciably and increase the reliability of the resulting scores. Combining responses to several related questions will increase the reliability of the composite score. The interpretability, or validity, of a composite score, however, depends on the extent of the common meaning in the component questions, and increased reliability of a composite score may come at the expense of decreased meaning in the composite.

Nominal scale reliability

Cohen (1960) proposed a statistic, kappa, to measure the degree of agreement between two raters, and Fleiss (1971) extended the use of kappa to any constant number of judges. Kappa is the increase in agreement among the judges over the degree of agreement to be expected by chance in relation to the maximum possible increment in agreement. For example, if two judges assigned responses randomly to one of five categories, agreement would be expected on 20 percent of the judgments. If the five categories differed in difficulty so that only a few re-

sponses were assigned to one category by any of the judges and correspondingly more were assigned to other categories, the chance agreement would be slightly higher than .20. If the proportion of judgments in agreement were .60 while the proportion of agreements expected by chance were .20, the increase over chance would .40. But since the maximum possible agreement is 1.00, the greatest possible increase over chance would .80, and kappa would be .40/.80 or .50.

Kappa requires no assumptions about category values or about the order of the categories; the categories are treated entirely nominally. Since the scorers in the present project were given descriptions of response categories and asked to assign responses to categories wholly on their substantive characteristics, without regard for judgments of merit or order, kappa is directly interpretable in terms of the scoring process used. It is the proportional improvement of the actual scoring process over a random process. Interpretation is less clear with respect to what an acceptable value for kappa should be in circumstances in which the questions of the present project might be used. The questions can be ordered in terms of kappa, however, indicating accurately their relative differences in the reliability of the scoring process. The magnitudes of the differences can be assessed in relation to the standard deviation of kappa and can be considered more realistic indicators than the intraclass correlation coefficient of differences in the reliability with which particular questions can be scored.

Values of kappa for the 32 most promising questions are shown in Table 1 with the other scorer reliability coefficients. The relationship between the intraclass correlation coefficient, based on a ratio of

variances, and kappa, based on an increase over chance in the proportion of consistent category assignments, is fairly high despite systematically lower values for kappa. For a few of the questions, however, the differences are appreciable. For those questions, most of the scorer disagreements involved adjacent categories and therefore comparatively small errors in the variance-based measure. With kappa, however, every disagreement between any two categories is equivalent to every other disagreement. To the extent that the order given to the categories is an accurate reflection of relative merit, kappa will underestimate scorer reliability. If the categories are only nominal distinctions without an interpretable order reflecting an underlying scale of merit, the intraclass correlation coefficient is an inappropriate measure that would differ from kappa in an unsystematic way. In a rough sense, then, kappa can be considered an underestimate of scorer reliability. While the intraclass correlation coefficient gives a higher value than kappa, it is not necessarily more accurate. The choice between the two measures depends on the confidence to be attributed to the relative order of the categories and to the equivalence of the distance between categories. The systematically higher values for the intraclass correlation coefficients imply some validity in the ordering of the categories. Disagreements among scorers are less likely when the categories are more widely separated in order.

Both forms of reliability coefficients refer to the degree of agreement between scorers assigning the same responses to categories. Variations in the students' responses that might be observed if the same students responded to the same questions at some different time or in different circumstances are not reflected in these reliability measures. The reliabilities of the responses

themselves, as opposed to the reliability with which they can be assigned to categories, have not been determined.

Difficulty

The early tendency of the faculty members devising the questions was to pose puzzles or to ask questions that required unusual subtlety of thought. Many of the questions were too difficult for most college students who had not had direct experience with the particular content of the questions. With successive trials, questions were revised to remove unnecessary obstacles to a successful response while retaining the requirement that the student exhibit the kind of thinking being tested. To a large extent, the final set of 32 questions can be managed with some success by students through a wide range of ability. An advantage of the open-ended format is that almost all students can give a response having some merit to most of the questions. Student reactions ranged from chagrin at their inability to produce completely adequate responses to enjoyment at the novelty and challenge of the questions.

Table 2 shows the distributions of responses for the 47 questions scored in the final administration. The four most difficult questions, as indicated by the percentages of students scoring in the top two categories were Questions 215, 321, 112, and 143. From 8 to 16 percent of the students produced answers to these questions that were scored in the top two categories. From 22 to 46 percent of the students scored in the bottom two categories of the same questions. In general, the response distributions were fairly broad and flat, indicating that most students have some success with every question and also indicating good discrimination if the distinctions among the categories are clear.

Relationships with aptitude

In Table 3 are shown the correlation coefficients among the 47 questions used in the third administration and with combined verbal and mathematics scores from the Scholastic Aptitude Test (SAT). The SAT was not administered as part of the study; scores were provided from the institutions' records. Since scaled scores were available for only 17 of the 47 questions, 3 of the 20 that had been scaled having been abandoned, the correlation coefficients were calculated using the ordered categories as though they were equal-interval scales. Although that assumption is not correct, the differences between coefficients based on scaled scores and category scores were slight.

The correlations between SAT scores and the 47 questions used in the final administration ranged from essentially zero to .52 with a median of .26. The following table shows the rough distribution of correlation coefficients between the 47 questions and SAT scores.

Value of Correlation Coefficients	Number of Correlation Coefficients
> .50	2
.40 - .49	4
.30 - .39	8
.20 - .29	20
.10 - .19	8
.00 - .09	5

The coefficients greater than .30 tended to be with questions measuring communication and analytic skills rather than with synthesizing and awareness questions. Questions in the humanities tended to have lower correlations with SAT than did questions in the other academic areas.

Interpretation of responses

The effects of errors in scores, whether in the responses themselves or in the scoring of the responses, are reduced if scores of a number of questions measuring the same competence are combined. For example, if six questions, all measuring the same competence and having intraclass reliabilities of .50, were combined to produce a single score, the reliability of the combined score would be .93. That combined score would be an accurate measure of competence for individual students. But the qualification that the questions all measure the same competence is severe. Questions such as those developed in the present project do not measure quite the same competence even when developed to be comparable and showing moderately high relationships with each other. The interpretation of a combined score refers to whatever is measured in common by all the questions pooled, a determination that is difficult to make but essential to any assessment that combines the scores of several questions. Identifying sources of variance that are demonstrably common to groups of questions rather than being presumed so because of similarity in the content or form of the question is a major problem.

When interest is on group rather than individual performance, the problems of reliability and interpretation are less severe than with individual scores but are still formidable. The information of interest is the way the members of a group of students distribute themselves among the various types of response, whether ordered or not. Errors in the responses themselves can be assumed to produce equal errors of exclusion and inclusion in the response categories, which will have a leveling effect on the response distributions. Fewer common responses and more uncommon responses will appear than would be the case if the responses reflected student competence accurately. Unless response

errors are quite large, the shape and location of the distribution of responses will not be seriously affected, and inferences about the collective performance of the group will be credible. Small groups of questions having some variance in common are sufficient for defensible conclusions about the competence of the group in terms of that common variance.

The spread of the responses over the 5 to 9 categories of the 32 selected questions was reasonably uniform, with most categories receiving a moderate number of responses, as shown in Table 2. Few categories other than the lowest in each question had fewer than 5 percent and only an occasional category accounted for more than 30 percent of the responses to a question. The easiest was Question 342, on which 148 responses were distributed over the five categories in percentages that were, in order of decreasing merit, 38, 22, 18, 11, and 11. Even though almost 40 percent of the students produced the highest category of response, more than 20 percent fell in the lowest two categories. The most difficult of the 32 was Question 112, for which 151 responses produced a distribution from high to low of 7, 7, 25, 15, 19, and 27 percent in its six categories. Almost half the responses were in the lowest two categories, but 14 percent made the top two. These questions were the extremes in terms of skewness of the distributions. More representative was Question 413, for which 150 responses were distributed across the six categories in percentages of 11, 23, 10, 18, 19, and 19.

Comparatively flat distributions such as these would be produced if the responses themselves were almost completely unreliable or if the scoring was unreliable. The scoring was clearly reliable enough to discriminate well among responses having different characteristics, and the nature of the responses themselves as well as the students' comments make highly unlikely the supposition

that random error was largely responsible for the flatness of the observed distributions. That the observed discriminations among types of response were realistic reflections of differences in student performance on individual questions seems quite clear. Identifying the precise nature of those differences in performance is another problem.

Questions within each of the four competences could be subgrouped according to apparent similarities in the ability required for the better responses. If the questions in these subgroups then showed higher relationships among themselves than with other questions, they could be treated as independent measures of a common competence. The results were ambiguous for several reasons. First, the scorer reliabilities, which were in general high enough for sensible interpretations of scores formed from several questions or of group scores for individual questions, were still low enough that the sizes of the correlation coefficients that could be expected were severely limited. Second, the response reliabilities, which would reduce the expected correlation coefficients still further, could not be calculated. Third, the sample sizes for any particular pair of questions varied, at times being zero and at other times being too small to provide a correlation coefficient in which much confidence could be placed. Nevertheless, the possibility of strong relationships appearing through the noise makes examination of the results worthwhile.

For example, Questions 315, 325, 332, and 342 all require the student to select specific aspects of the information presented and synthesize them into a general statement. Although each represents a different area of content, similarity in their presumed task should produce moderate correlations among them. The correlation coefficients based on scaled scores are shown in Table 4.

Of the six correlation coefficients among the four questions, two were unobserved because none of the sets of six questions completed by the students included those two pairs. A third was essentially zero, and the other three were low but positive. If only the effects of scorer reliability are considered, the three positive coefficients corrected for scorer error would be on the order of .40 to .50. Correction for error in the responses themselves would raise the true correlation coefficients somewhat higher still.

Examination of the values of gamma, however, a measure of association based on ordered categories, raises doubts about the relationship of Question 315 to the other three questions. Since gamma (Goodman & Kruskal, 1954; Hays, 1963) assumes only an ordered relationship among the categories of each question without the requirement that the intervals between categories be equal, as does the correlation coefficient, it is a more appropriate index of relationship for the present data. In the absence of data on two of the six coefficients that would show the relationships among the four questions, a tenuous case can be made for three of the four questions being mutually related.

The single clearcut relationship in Table 4 based on a sizeable sample is that between Questions 325 and 342. The source of that relationship is largely in the tendency for students who give high-quality responses to one question to give high-quality responses to the other question as well. The top two categories of Question 342 require the student to select related elements from a given list and from them formulate a general statement. The top two categories of Question 325 also require the student to formulate a general principle from elements he or she selects from the material presented, but the nature of the material in the two questions differs appreciably.

Question 342 presents a list of verbal statements about characteristics of the federal government while Question 325 presents a symbolic figure, but the tasks required are similar. The poorer responses of the two questions are unrelated, suggesting that failure to find and describe an integrating principle may not arise from a common deficiency. This observation anticipates one to be discussed more fully in a later section--that the better and poorer responses may not reflect higher and lower positions on a common dimension.

Question 332 showed a weak but positive relationship with Question 325. The presented material for Question 332 is a graph from which students are to select information to be integrated into a generalization. The low and statistically not significant relationship with Question 325 may be due to low reliability in the scoring process. Scorer agreement on the sample of 20 responses drawn from the final administration was poor. Yet the two earlier trials both showed fairly good scorer reliability for the same question with very minor modifications, so the reliability shown in Table 3 is likely to be an underestimate.

The most sensible conclusion from Table 4 is that of four questions constructed so that they seem to be measuring a common competence, two seem moderately well related to each other, one seems unrelated to at least two of the other three, and the fourth is equivocal in its meaning. Additional data might show Question 332 to form a triad with Questions 325 and 342, and related questions might then be developed to substantiate their interpretation as measures of an ability to formulate a general statement from specifics given in a variety of forms.

If the two empirically related questions, 325 and 342, are taken as a starting point and a search is made for other questions related to them, several suggestive results are found. Question 234 shows a statistically significant gamma ($p<.01$) with Question 325, but data are not available to compare it with Question 342. It requires students to compare several different sources of data and select the one most likely to permit defensible inferences. The best responses of Question 234 relate possible inferences to the information provided, a task closely related to the generalization from specifics required by Questions 325 and 342.

Two other questions, both measuring communication ability, showed unanticipated relationships with both Questions 325 and 342. Question 133 showed gammas of .33 and .44 with Questions 325 and 342 respectively, both statistically significant ($p<.01$). Question 133 presents students with information in a set of jargonladen statements and asks them to restate the information concisely in everyday language. The top two categories require the response to indicate "an integration of the information rather than a paraphrasing of it." The possible relationship between this requirement and the generalization or synthesizing of Questions 325 and 342 seems clear.

The other communication question related to the two questions measuring synthesizing ability is Question 142. It presents students with a brief statement of a concept and asks them to expand it and apply it in a setting different from the illustration given them. Both top categories require "a general statement of the concept," as with Question 133, suggesting a task having some relationship to the synthesizing of Ques-

tions are respectively .32 and .40, but the numbers of cases on which they are based are both less than 30 and too small for the gammas to be statistically significant. A significant relationship ($p<.02$) does appear between Questions 142 and 234, the other question found to be related to one of the synthesizing questions. The relationships among these five questions--the two clearly related synthesizing questions plus the three additional questions related to one or both of them are shown in Table 5.

Question 234 was designed to assess analytic thinking in the belief that the students would be required to identify the distinctions among five sources of data presented to them. Among 14 other questions with which it was compared, the only relationships of any magnitude were the two shown in Table 5, suggesting that the identification of inferences possible from different types of data may be more closely related to synthesizing than to analytic ability.

The five questions of Table 5 make a promising candidate for a cluster of related questions that could be used to assess some aspects of the synthesizing or generalizing ability of a group of students. All five, but particularly the two communication questions, might be revised to bring them more closely into focus on a common ability. Until the observed relationships can be confirmed and the unobserved relationships examined, however, they offer only a promise.

Discrimination Among Groups

Year in college

If the questions do reflect the kinds of academic performance that colleges and universities promote in their students, upper-division students would be expected to score higher than lower-division students, and seniors appreciably better than freshmen. In general that was not the case. Either most of the questions were not measuring what they appeared to be or those capabilities were not being developed in the 15 participating colleges and universities. The project was not a rigorous test of college or program effects, though, because the samples were small and not representative of the student bodies or programs at the various institutions. An average of 68 lower-division and 83 upper-division students responded to each question, numbers too small to consider representative of the student bodies of 15 institutions.

Two of the 32 selected questions did discriminate fairly sharply between upper-division and lower-division students, and a third showed a less clear-cut difference. All three were questions on awareness, and all three showed upper-division students to be superior. Of the two questions that clearly discriminated between upper- and lower-division students, Question 412 showed a uniform increase in scores from freshman to senior years, while Question 441 showed a sharp break between the sophomore and junior years and Question 424 showed a similar but weak split at the same point. Since the questions were administered near the end of the academic year, that distinction is between students

who had completed one or two years of college and those who had completed three or more years.

The differences in the distributions of the responses across the 6 or 7 categories were tested for statistical significance using the Kolmogorov-Smirnov test (Walker & Lev, 1953, pp. 426-428). The two clearly discriminating questions showed differences significant below the .01 level while the third question showed a difference slightly below the .10 level of significance. The distributions across response categories grouped into those judged by faculty to be High, Medium, and Low in quality, are shown in Table 6.

Curricular pattern

One of the purposes of the project, if demonstrably effective measures could be developed, was to compare the performances of students who had been through different types of general education programs. The students were classified into one of three curricular patterns: (1) general education requirements were satisfied by completing a minimum number of units in humanities, social sciences, and natural sciences; (2) a prescribed group of general education courses were completed; (3) the students followed individualized programs of study. About 150 students were classified in each of the last two categories. The need to examine results question by question rather than through the pooling of questions, however, made curricular comparisons infeasible. The numbers of students in the last two curricular categories were only 18 or 20 per question.

Institutional selectivity

The wide variability across institutions in the academic aptitude of their students might account for the failure of the questions to discriminate across year in college if the sample of students drawn from each institution was not representative. Freshmen at the most selective institutions can be expected to perform better than many seniors at less selective institutions, and if sampling inadequacies produced proportionately more seniors than freshmen at nonselective institutions and the reverse at selective institutions, differences associated with year in college might well disappear. Many of the questions were moderately to highly correlated with aptitude, and the combination of high aptitude and a selective college environment could be expected to affect performance on the questions. To examine possible effects of institutional selectivity, the 15 institutions were grouped according to the average aptitude of their entering freshmen. A high group of four institutions had average SAT scores for entering freshmen of about 1120 or higher. The low group of seven institutions had corresponding scores of about 1000 and below. The average scores of the other four institutions ranged from about 1020 to 1080. The response distributions of students in the four most selective and seven least selective institutions for five questions in which the distinctions were fairly large are shown in Table 7. All favored the students at selective institutions. The remaining 27 questions among the 32 selected showed essentially no difference between the two groups or a moderate difference in favor of the selective institutions.

For each of the pairs of distributions in Table 7, the selective institutions have proportionately more students in the better response categories

and proportionately fewer in the poorer categories. The Kolmogorov-Smirnov test of differences in cumulative percentage distributions shows levels of statistical significance less than .10 for two of these five pairs of distributions, less than .05 for one pair, and less than .01 for two pairs. While the interpretation of statistical significance tests for data from two groups formed somewhat haphazardly is questionable, these differences are greater than those likely to be produced by sampling error.

Correlations of these five questions with SAT scores ranged from 24 to .49, only slightly higher than most of the correlations of other questions with SAT. Twenty of the 32 questions had correlations in that range or higher. The association of those questions with institutional selectivity is therefore not likely to be accounted for wholly by the academic aptitude of the students responding.

Shifts over time in distributions of responses to questions such as these, or failure to shift, or comparisons of groups of students engaged in different kinds of academic programs can reveal the nature of the learning that is or is not occurring. The present questions do not yet permit clear enough interpretations for purposes such as those, but their potential value seems appreciable.

Faculty Standards

One benefit of having a group of explicit but divergent measures of student performance is that they can be used to define faculty standards. If the measures are explicit, desirable standards for students can be made explicit, at least in principle. In an attempt to define faculty standards in terms of the developed measures, faculty members were asked to judge "how a properly educated group of college seniors would distribute themselves" across the different levels of response of a number of questions.

These judgments were constrained in several ways expected to make their results interpretable. Since the faculty members were asked to define the levels of learning to be expected in a properly educated group of seniors, their judgments should not have been colored by jaundiced views of current poorly prepared students. Applying their judgments to a group of seniors rather than to a single ideally educated senior should allow for the realistic variation in performance to be expected even in well educated groups of students. Finally, asking for judgments with respect to specific kinds of performance should give clear definition to whatever standards the faculty members set. The results were interesting but less definitive than had been hoped.

Eight questions, two for each of the four competences were selected for judgments. Three responses to each--one High, one Medium, and one Low in quality--were described in terms of the characteristics expected of responses at that level and were illustrated with representative responses. The faculty members' task was to judge what proportion of their hypothetical group of properly educated seniors would give each type of response.

No constraint was placed on the form of the distribution. The entire group could be judged in one category if that seemed appropriate. In addition to the three described responses, four additional categories of response were provided--one above the best response illustrated, one below the poorest response, and two in between the three illustrated responses--giving a total of seven categories.

The results in Table 8, comparing the judgments of 48 faculty members with actual student percentages, show the students falling generally below faculty expectations but with a number of exceptions. On four of the questions, 112, 143, 244, and 436, student performance fell below faculty expectations while on one question, 223, it exceeded them. On one question, 322, performance and expectation were quite close. On the other two the results were mixed. On 315, more students than expected scored in the High and the Low categories. On 414 that pattern was reversed, fewer students than expected falling in both the High and Low categories.

Attempting to define faculty standards and compare student performance with them from this small set of data would be futile. Yet if the different types of responses were clearly indicative of the desired kinds of performance, and if similar kinds of responses to different questions were clearly related to each other, faculty expectations could be made explicit. For example, Questions 223 and 244 are both intended to measure analytic thinking, one with content in English and one with content in political science. Both require superior responses to identify the key elements in the presented information, and both elicit two kinds of poor responses, one in which the student makes gratuitous assumptions

beyond the information given and one in which the student does no more than give back the presented information. Faculty expectations were quite similar for the two questions, but students regardless of their field of study found the political science question more difficult.

The possibilities in identifying common faculty expectations across disciplines and comparing student performance on those common abilities with content specific to either discipline have great appeal. The generalization of learning across disciplines could be assessed and fostered. Failure to find generalization, as when students perform a similar task well in one discipline but poorly in another, would identify specific areas to which instruction could be directed. Examining the reasons for dissimilar kinds of performance in different fields would probably lead to greater clarity of instructional purpose in both fields as explanations are sought for the differences.

Other implications of the kind of data presented in Table 8 can be easily imagined if the meanings of the different responses to the questions are clear. Differences in performance across student groups--curricular groups, for example--would be informative. So too would be differences between student performance and faculty objectives, as is suggested in Table 8. Clarity of teaching objectives would probably be greatly improved by faculty members first deciding individually what their expectations with respect to certain kinds of performance ought to be and then discussing discrepancies between their own expectations and those of other faculty members.

Each activity suggested above would require, first, demonstrations of validity in each of the questions and their different responses, and second, a number of questions closely enough related in the kinds of performance called for to provide mutual confirmation of the inferences to which they lead. Those two requirements which are only slightly different aspects of validity, were sought without success in the present project. Yet both seem feasible with further work, and the consequences of reaching them are appreciable.

Levels of Performance

The process of placing response categories on a scale of merit based on faculty judgments was not entirely satisfactory. Faculty members disagreed markedly with respect to the relative merit of a number of categories. Yet closer examination showed the disagreement to occur predominantly among the middle-ranked categories. For almost every question, the top one or two categories stood out as clearly superior to the other types of response. The faculty members were also moderately consistent in their judgments of the responses assigned to the poorest categories, though not as much so as with the best responses. This pattern of faculty judgments is understandable and would be expected if students in the middle range of performance fall short because of one of several different kinds of deficiency.

In communication, for example, students may fall short of a top response because of a style that was verbose or ostentatious, because of poor diction and inelegance, or because of inaccuracy or incompleteness. Deficiencies of any of these kinds keep students out of the top category and put them into one of several intermediate categories. While the relative merits of these intermediate categories are not clearly differentiated, their qualitative distinctions are.

A similar set of distinctions occurred with the other three competences. In analytic thinking, the intermediate responses fell short of the top categories because of overgeneralizing, undergeneralizing, or failure to perceive or to select the critical points. In synthesizing, the intermediate responses reflected description without integra-

tion, erroneous perceptions or interpretations of the given material, and overgeneralization. The intermediate responses to the questions measuring awareness showed insufficient grasp of (1) the scope of an issue, (2) the relationships among the several elements of an issue, or (3) the implications of an issue or its metaphorical meaning.

If these types of intermediate-level responses can be confirmed and validated more fully, questions such as the ones developed in the present project would have diagnostic value. The diagnoses would still apply to groups of students rather than to individuals, but deficiencies in groups of students can suggest ways to adjust instruction to correct the deficiencies. Or enough additional questions that focus on a particular deficiency can be administered to the group to identify those students who tend toward particular kinds of errors. The most reasonable approach would probably be to spend some instructional time with the entire group of students on the types of deficiencies most commonly found. Devising quizzes or mid-term examinations that are based on current course material but that also require thought processes similar to those in which the students were most deficient should not be difficult.

The different types of deficiencies and their relative rates of occurrence in different fields and at different levels should also aid in curricular planning. Standards of performance can be set by faculty members collectively for a department or division or individually for a course, and student performance with respect to those standards can then be checked and curricula adjusted appropriately. The particular merit

of questions of the type here discussed is that they deal with kinds of student performance that are broader than learning related to the content of specific courses and that therefore have implications beyond a particular course or department. As must be repeated throughout this report, however, the meanings of the present set of questions, or the implications of their different responses, are not yet well enough confirmed to permit them to be used with confidence.

Implications

The development of open-ended questions that can be answered in a comparatively short time and that produce qualitatively different types of response that reflect different levels of academic performance is clearly feasible. Scoring of those questions, or classifying them into distinct categories, can also be done reliably and at a manageable cost in faculty time. Neither the development of the questions or their scoring categories is simple, but both tasks can be beneficial in sharpening faculty members' definitions of the kinds of learning they are trying to produce and of the kinds of student performance that would demonstrate the desired learning.

Several institutions not in the original group tried out the questions and the scoring process. The major criticism centered around the complexity of some of the scoring procedures. A second reservation concerned the difficulty of the questions for students in nonselective institutions. Both these concerns can be met with further work on new and modified questions to give them the desired characteristics. And that process itself is informative for the faculty members who try to devise questions that will elicit unambiguous indicators of the desired kinds of performance.

Two problems remain with respect to the present set of questions. First, the validity, interpretability, or meaning of the various response categories must be determined with greater clarity. Second, a sufficient number of questions providing responses that can be interpreted similarly must be developed to permit inferences about the performance of groups of students to be made with some confidence.

Both these problems seem tractable. Their resolution would lead to better understanding of the kinds of thought processes that underlie good academic performance, sharper definition of standards of academic performance, clearer specification of kinds of learning that cut across disciplinary boundaries (e.g., analytic processes in engineering and English), and manageable procedures for assessing the performance of groups of students and therefore the effectiveness of educational programs.

References

Bell, D. The reforming of general education: The Columbia College experience in its natural setting. New York: Columbia University Press, 1966.

Blackburn, R., Armstrong, E., Conrad, C., Didham, J., & McKune, T. Changing practices in undergraduate education. Berkeley: Carnegie Council on Policy Studies in Higher Education, 1976.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Fleiss, J. L. Statistical methods for rates and proportions. New York: Wiley, 1973.

Frederiksen, N., & Ward, W. C. Measures for the study of creativity in scientific problem-solving. Applied Psychological Measurement, 1978, 2, 1-24.

Glaser, R., & Resnick, L. B. Instructional psychology. In P. H. Mussen & M. R. Rosenzweig (Eds.), Annual review of psychology (Vol. 23). Palo Alto, Calif.: Annual Reviews, Inc., 1972.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Hook, S. General education: The minimum indispensables. In S. Hook, P. Kurtz, & M. Todorovich (Eds.), The philosophy of the curriculum: The need for general education. Buffalo, N. Y.: Prometheus Books, 1975.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Walker, H., & Lev, J. Statistical inference. New York: Henry Holt, 1953.

Warren, J. R. Describing college graduates in 87 phrases or less. Findings, 1976, 3 (2), 5-8.

Winer, B. J. Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill, 1972.

Table 1

Scorer Reliabilities for 32 Questions

	Question	ICC Unscaled[a]	ICC Scaled[b]	Kappa
111	Oil Refining	.89		.59
112	Verbiants	.90		.58
113	Equation Translation	.70		.32
114	Schrodinger	.40	.48	.21
122	Line Selection	.48		.16
124	Paragraph Style	.49	.44	.22
133	Life on Pluto	.52	.38	.29
142	Historical Thinking	.46	.40	.20
143	Nature of Government	.75		.31
144	Approaches to History	.50		.16
215	Hybrid Grasses	.58	.30	.38
222	Socrates	.57	.51	.24
223	O'Connor Study	.14		.08
224	John Stuart Mill	.84		.53
231	Chimpomat	.63		.19
232	US-Russia	.66		.45
234	Capital Punishment	.89	.89	.48
244	Oppenheimer	.62		.14
245	Roman Conquest	.48		.27
315	Phytoplankton	.66	.66	.39
321	Creative Endeavor	.47		.35
322	Litton Industries	.44	.47	.35

Question		ICC Unscaled[a]	ICC Scaled[b]	Kappa
325	Yin-Yang	.54	.60	.19
332	Suicide	.28	.21	.21
342	Government Strengths	.54	.50	.19
412	Earth's Rotation	.57		.18
413	Ephemeral Insects	.62		.23
424	Haiku	.66		.38
431	George Herbert	.42	.10	.42
436	Canada	.71	.56	.56
441	Presidential Candidate	.68	.62	.48
444	Poor and Mean	.42		.15

[a]Intraclass correlation coeffient computed from consecutive numbers assigned to the ordered categories.

[b]Intraclass correlation coefficient computed from scaled values assigned to the categories.

Note. The first digit of the question number indicates the competence measured while the second indicates the question's field of content according to the following codes:

Competence	Content
1 Communication	1 Natural Sciences
2 Analysis	2 Humanities
3 Synthesis	3 Social Sciences
4 Awareness	4 History/Political Science

Table 2

Response Distributions in Percentages

Question	Number of Respondents	Response Categories[a]								
		1	2	3	4	5	6	7	8	9
111	149	13	16	16	20	20	16			
112	151	7	7	25	15	18	27			
113	146	34	15	20	26	5				
114	147	22	5	4	21	24	7	15	3	
121	155	37	4	41	17					
122	150	15	35	20	17	12				
124	148	14	34	23	18	11				
132	147	12	6	2	3	23	2	35	16	
133	151	25	15	15	5	22	7	10		
142	152	20	18	3	32	9	16			
143	151	6	10	8	17	3	6	9	18	23
144	147	24	24	8	16	18	10			
211	153	26	18	28	7	14	8			
213	152	15	14	18	12	18	18	4		
215	157	4	4	32	27	20	12			
216	146	23	12	19	8	38				
222	151	18	11	13	11	24	24			
223	150	33	13	21	33	1				
224	155	18	10	27	19	26				
231	153	5	14	21	14	20	14	12		
232	148	18	24	21	18	7	12			
234	160	12	24	14	18	21	2	9		
244	153	13	14	16	14	27	16			
245	154	17	18	29	12	6	18	1		
313	163	17	28	23	22	10				
315	182	15	25	18	30	8	4			
321	152	3	7	22	30	20	16			
322	149	17	9	23	28	13	11			

Question	Number of Respondents	Response Categories[a]								
		1	2	3	4	5	6	7	8	9
323	178	10	11	22	6	14	33	3		
324	153	46	6	5	28	6	3	6		
325	157	22	31	13	22	10	2			
326	151	12	35	35	4	12	2			
332	149	22	31	12	32	3	1			
342	148	38	22	18	11	11				
344	156	22	3	40	15	8	10	3		
411	151	20	25	22	11	15	8			
412	148	11	14	38	15	12	11			
413	150	11	23	10	18	19	19			
414	151	13	40	5	34	6		3		
424	150	7	25	19	11	31	7			
431	153	41	16	23	2	18				
433	150	26	25	1	15	24	9			
434	144	15	3	19	59	4				
436	154	8	30	3	28	13	14	4	1	
441	143	15	18	27	7	6	28			
443	148	7	31	20	14	17	12			
444	151	9	17	28	22	10	15			

[a]The best responses are in Category 1 with quality decreasing to Category 9.

Table 3

Correlation Coefficients Among the 47 Questions of the Third Administration and with SAT

Question Number	SAT	111	112	113	114	121	122	124	132	133	142	143	144	211	213	215	216	222	223	224	231	232	234	244	245
SAT	766	.52	.32	.28	.13	.22	.33	.27	.33	.08	.40	.35	.26	.21	.27	.26	.34	.36	.19	.49	.31	.10	.24	.30	.07
111	87	149			-.09	.15		.13				.03		.21			-.03	.06		-.01				.27	
112	92		151			.23			.22		.07		.07		.07			.44			-.22		.06		-.21
113	97			146	.41				-.04	.24		.11				.33	-.02				.10	.11		-.03	
114	89	68		72	147			.13		.21				.17		.29	.12			.33		.30			
121	98	75	75			155					.10	.10		.09	-.12			.18					-.19	.10	
122	88						150	.39		.07	.35		.11						.05	.26		.20	.07		-.18
124	80	69			70		74	148			.11			-.04			.19		.22	-.00			.20		
132	92		73	70					147			.24	.27		.11	.18					.02			.01	.13
133	97			74	74		75			151			.18			.10	-.14		.17			.18			.17
142	90		76			78	73	73			152				-.16			.12	.23	.35			.29		
143	99	75		71		76			72			151		.10		.27		.12			.22			.40	
144	88		72				73		73	72			147		.12				.21		-.30	-.06			-.27
211	91	49			23	24		22				25		153			-.04	.18		.30				.12	
213	90		49			25			25		28		29		152			-.07			.13		-.23		.16
215	92			51	25				25	26		25				157	.10				.01	.21		.13	
216	78	21		26	45			20		26				66		74	146			.04		-.09			
222	98	25	24			49					25	25		76	66			151					.20	-.03	
223	97						49	26		22	26		22						150	.18		.17	.09		-.13
224	86	24			22		28	50			28			74			66		73	155			.16		
231	96		25	25					50			25	26		74	78					153			.26	.13
232	90			25	24		23			47			22			71	70		75			148			-.05
234	97		23			24	27	27			51				63			54	73	80			160		
244	99	25		25		24			25			50		74		77		74			78			153	
245	97		25				25		25	25			50		81				73		72	73			154

Note. Correlation coefficients are in the upper half of the table; numbers of cases on which they are based are in the lower half; the numbers of students responding to each question are in the diagonal.

Table 3

Correlation Coefficients Among the 47 Questions of the Third Administration and with SAT

Question Number	SAT	111	112	113	114	121	122	124	132	133	142	143	144	211	213	215	216	222	223	224	231	232	234	244	245
313	106	52			26	29		26				29		52			22	27		25				25	
315	107		55			29			25		30		24		54			29			25		53		24
321	91			43	22				23	22		24				53	26				27	24		26	
322	86	23		20	47			26		21				25		26	48			25		24			
323	115	26	28			53					28	26		25	26			50					50	23	
324	99		23				23		23	23			43		32				26		25	26			52
325	103						47	23		24	22		22		10				51	28		26	28		28
326	92	24			26		22	48			22			25			22		23	51			26		
332	91		25	23					48			24	24		25	26					51			25	24
342	89			20	22		20			42			19			25	25	10	25			48			27
344	105	28		24		28			24			53		23		25		24			25			46	
411	84	43			22	23		23				23		51			23	24		26				24	
412	90		46			23			25		23		24		24						23		25		22
413	97			48	25				23	26		23				52	24				27	23		27	
414	87	24		26	48			25		26				26		26	47			26		23			
424	93	20	22			45					23	22		26				26					25	25	
431	97						47	22		25	22		25						50	25		26	27		29
433	83	25			24		22	48			22			26			23		24	51			27		
434	88		23	22					46			22	24		23	26					50			27	22
436	96			27	25		25			51			25		14	26	27	13	25			48			28
441	87		22			23	22	22			45								23	22			48		
443	93	20		20		22			21			44		26		26		25			27			52	
444	96		24				24		25	24			48		27				26		23	26			51

Table 3

Correlation Coefficients Among the 47 Questions of the Third Administration and with SAT

Question Number	313	315	321	322	323	324	325	326	332	342	344	411	412	413	414	424	431	433	434	436	441	443	444
SAT	.13	.29	.14	.22	.27	.41	.16	.27	.03	.21	.25	.28	.47	.24	.15	.25	-.05	.27	.12	.06	.50	.23	.29
111	.16			.42	.19			.00			-.05	.05			.30	.21		-.19				.33	
112		.24			.17	.12			.31				.20			.02			-.20		.01		-.08
113			.02	.32					.43	.19	.54			.09	.18				-.38	.51		-.16	
114	.04		-.17	-.01				-.15		-.14		.19		.08	.10			-.15		.50			
121	-.16	.56			-.03						-.27	.41	.21			-.15					-.20	-.06	
122						.03	.35	.18		-.06							.12	.40		-.22	.25		-.08
124	.29			.45			.14	.22				-.22			.13		.28	.12			-.13		
132		-.03	.32			.28			.08		-.22		.31	.08					-.15			.26	.27
133			-.12	-.02		.17	.47			.28				.14	.18		-.08			-.06			.02
142		.15			.21		.40	.38					.19			-.10	.16	.51			.38		
143	-.26		.05		.34				.12		.00	.45		.32		.03			.29			.17	
144		.18				.07	-.06		.22	.17			.24				.51		-.01	.19			.32
211	-.02			.60	.24			.29			.07	.14			.08	-.13		.16				-.08	
213		.17			.06	.07	.45		-.10				-.22						-.17	.45			.35
215			.09	.15					.20	.13	.20			.13	.16				-.04	-.04		.26	
216	-.01		.31	.21				-.01		-.01		.26		-.02	.29			.44		.18			
222	.19	.42			.24					.08	-.12	.04				.01				.38		.09	
223						.01	-.08	-.09		-.01							.27	.04		-.03	-.23		.18
224	.34			.05			.13	.02				.03			-.19		.03	-.22			.43		
231		.01	-.02			-.07			.05		-.12		.14	.13					-.10			-.10	.54
232			.06	.19		.06	.03			.32				-.04	.09		-.18			-.19			.15
234		.19			-.04		.30	.03					-.30			-.15	-.17	-.26			.12		
244	-.38		.12		.02				-.13		-.24	.32		.10		-.02			.01			.19	
245		.37				.09	.00		.02	.42			.28				.42		-.02	.10			.08

Table 3

Correlation Coefficients Among the 47 Questions of the Third Administration and with SAT

Question Number	313	315	321	322	323	324	325	326	332	342	344	411	412	413	414	424	431	433	434	436	441	443	444
313	163			.15	-.14			.01			.00	-.28			-.10	.04		.26				.16	
315		182			-.03	.05			.14				.19			-.11			-.21		.37		-.16
321			152	.22					.21	.22	.25			.07	-.12				.02	.21		-.30	
322	77		70	149				.24		.02		.05		.03	.01			.10		.14			
323	76	101			178						.08	.17	-.29			.05					.09	.20	
324		73				153	.00		.08	-.15			.01				-.20		-.04				.22
325						79	157	.19		.29							.40	-.11			.03		-.06
326	77			77			73	151				-.11			-.14		.10	.25			.35		
332		75	73			73			149		.28		.26	-.46					.01			-.08	.04
342			71	70		74	75			148				-.31	-.01		.34			-.34			-.08
344	80		74		72				73		156	-.07		.00		.19			-.16			-.04	
411	53			25	24			25			27	151			-.02	.25		.27				-.02	
412		73			48	25			25				148			.17			.08		.36		.20
413			47	23					24	24	24			150	.18				.05	.08		-.02	
414	24		23	47				24		24		72		75	151			.07		.26			
424	27	47			71						28	74	74			150					.06	.15	
431						24	49	25		23							153	.03		.10	.08		-.09
433	24			24			25	49				73			74		74	150			.20		
434		25	23			25			48		23		72	72					144			.10	.10
436			26	23						28				75	75		57			154			.33
441		47			47		24	24					74			74	69	69			143		
443	28		22		24				22		49	74		70		73			69			148	
444		25				49	24		25	23			73				78		72	53			151

Table 4

Relationships Among Four Synthesis Questions

	Q315	Q325	Q332	Q342
Q315	.66 (20)	-.02 (21)	.14 (75)	--
Q325	-.09 (21)	.60 (20)	.17 (21)	.29 (75)
Q332	-.07 (96)	.19 (21)	.21 (20)	--
Q342	--	.39 (75)	--	.50 (20)

Notes. Figures above the diagonal are correlation coefficients based on data from the final administration to approximately 1,200 students.

Figures in the diagonal are scorer reliabilities based on 20 randomly selected responses and four independent scorers.

Figures below the diagonal are gamma coefficients based on combined data from all three trials except when the questions had been modified for a later trial.

Figures in parentheses are the numbers on which the coefficients are based.

Dashes indicate pairs of questions for which joint responses were not obtained.

Table 5

Unanticipated Relationships Among Five Questions Bearing on Synthesis

	Q133	Q142	Q234	Q325	Q342
Q133	.38 (20)	--	--	.47 (24)	.27 (42)
Q142	--	.40 (20)	.29 (51)	.40 (22)	--
Q234	--	.30 (51)	.89 (20)	.30 (28)	--
Q325	.33 (71)	.32 (22)	.45 (28)	.60 (20)	.29 (75)
Q342	.44 (42)	.40 (28)	--	.39 (75)	.50 (20)

Note. See notes for Table 4.

Table 6

Questions Discriminating Across Year in College

	Year in College	N	Percentage Distributions Low	Med	High
Q412	Lower Div.	72	51.5	34.7	13.9
	Upper Div.	76	23.6	42.1	34.2
Q441	Lower Div.	65	52.3	24.6	23.1
	Upper Div.	78	30.8	28.2	41.0
Q424	Lower Div.	78	55.2	21.8	23.0
	Upper Div.	72	43.1	15.3	41.6

Table 7

Response Distributions of Questions Discriminating Between Selective and Nonselective Institutions (Percentages)

Question Number		N	Response Categories[a] 1	2	3	4	5	6
112	Selective	39	15.4	7.7	38.5	10.3	7.7	20.5
	Nonselective	74	4.1	9.5	14.9	20.3	25.7	25.7
142	Selective	38	26.3	13.2	5.3	47.4	2.6	5.3
	Nonselective	80	20.0	17.5	2.5	27.5	13.8	18.8
224	Selective	42	31.0	11.9	28.6	19.0	9.5	
	Nonselective	80	17.5	11.3	22.5	15.0	33.8	
234	Selective	44	20.5	34.1	18.2	2.3	15.9	9.1
	Nonselective	77	5.2	22.1	11.7	29.9	20.8	10.4
315	Selective	43	23.3	37.2	7.0	32.6	0.0	
	Nonselective	92	14.1	16.3	19.6	32.6	4.3	

[a]The best responses are in Category 1 with quality decreasing to Category 6.

Table 8

Faculty Expectations and Student Performance
(Faculty N = 48)

Question Number		Percentages of Students Giving Each Response High	Medium	Low
112	Faculty Judgments	23	48	28
	Actual Responses	7	67	27
143	Faculty Judgments	21	57	20
	Actual Responses	16	49	33
223	Faculty Judgments	29	51	22
	Actual Responses	33	67	1
244	Faculty Judgments	30	43	27
	Actual Responses	13	44	42
315	Faculty Judgments	28	52	20
	Actual Responses	40	18	42
322	Faculty Judgments	18	61	23
	Actual Responses	17	58	24
414	Faculty Judgments	29	52	19
	Actual Responses	13	79	9
436	Faculty Judgments	40	42	18
	Actual Responses	8	51	30

Appendix A

Definitions of Competences

Communication

High: Students high in this ability express themselves clearly and precisely, without unnecessary words, stating their ideas vividly but smoothly, economically but with elegance. They can shift easily from one mode of expression to another, from direct exposition to metaphor, from symbolic or graphic statements as in mathematics, science, or engineering to verbal expressions and back again. Their communication flows easily and effectively whatever its form.

Middle: Students at the middle level in this competence write understandably but without the force, clarity, or economy of expression of the better students. Their choices of words are not always precise and their verbal constructions are occasionally awkward. They manage straight verbal exposition acceptably but are clumsy with simile and rarely venture to metaphor. They shift reasonably well between verbal statements and nonverbal, graphic, or symbolic systems but without easy fluency.

Low: Students low in communication ability write statements that are often difficult to follow, unclear, or ambiguous. Words are poorly chosen and may be inaccurate. The structure of their sentences and paragraphs may be a hindrance rather than an aid to understanding. They have difficulty translating accurately between verbal and symbolic statements, often misinterpreting information presented graphically or symbolically. Accuracy of expression is difficult for them whatever form they use.

Analytic skill

High: Students high in analytic skill are quick to identify the essential components of ideas, events, problems, and processes. They distinguish the important from the unimportant, fact from conjecture, and make other distinctions imaginatively but usefully. They pick out quickly deficiencies or inconsistencies in statements or positions. They are realistically skeptical or critical but not destructively so.

Middle: Middle-level students find the essence of a problem or set of ideas through methodical effort rather than quick perception. They make useful distinctions among objects or ideas but not those that are unusual or imaginative. They find the obvious deficiencies in ideas or situations they are presented and effectively follow standard or accepted procedures to identify and solve problems. They are appropriately skeptical or critical but may miss a subtle inconsistency or fail to see an unusual but available resolution of an apparent problem.

Low: Students low in analytic skill have difficulty getting beneath the surface of a problem, idea, or a situation. They work with what they are presented, rarely dissecting it to examine the source or nature of a problem. They have difficulty distinguishing the essential from the unessential or making other discriminations that would simplify their task or lead to more effective solutions to a problem. They tend to be either uncritical, accepting what they are presented without question, or blindly critical, raising questions or objections that are neither well-founded nor useful.

Synthesizing skill

High: Students with a high degree of synthesizing skill are adept at organizing apparently unrelated ideas into a common frame of reference. They find unifying themes or common threads that let them deal effectively with diffuse bodies of information. Facility at constructing a general framework to accommodate whatever information they have lets them move easily beyond that information, applying its principles in new situations, generalizing usefully to other bodies of information, or extrapolating from given to unknown situations. In short, they synthesize what they know into more general intellectual structures that they can bring to bear on new problems.

Middle: Students at an average level of synthesizing make connections between ideas when they are apparent or have been pointed out. Having seen the connections, they can then apply them and may also make similar connections with similar ideas. They are not likely spontaneously to form connections among ideas or to make useful generalizations much beyond the given material. They grasp most of the material presented them either in the classroom or textbooks but may have difficulty with highly complex and abstract systems of concepts.

Low: The students low in this skill have difficulty relating ideas into higher order abstractions. They may grasp what is presented to them at the moment but do not readily integrate the current ideas with what has gone before or appreciate their ramifications into other situations. Their generalizations are often faulty, appearing to result from a quick focus on one or two salient points rather than from more complete understanding.

Awareness, sensitivity

High: The students high in awareness or sensitivity see the broad ramifications and implications of ideas, events, and issues. They are not bound to the particular or immediate perspectives from which an issue is presented to them. They are aware of the cultural roots of value systems and of the impact of values, including those of value systems other than their own, on human affairs. They appreciate the historical development of human culture and the relationship of present values to those of the past.

Middle: Students in the middle range of awareness and sensitivity understand the broad implications of ideas, events, and issues as they are pointed out to them, but carry them no further on their own. They understand in general that differences in value systems affect the ways people perceive and respond to a given situation but do not readily see the operation of a particular set of values in a particular instance. They speak of issues like cultural diversity and historical dependence but in phrases too trite to indicate much depth of understanding. Their awarenesses tend to be superficial.

Low: Students who are low in awareness lack both breadth and depth of perspective. Their perceptions tend to be culture-bound, unchanged by exposure to the ideas or values of different times or places. While they may recite an idea from a point of view other than their own, they cannot seriously examine the possibility of its validity.

Appendix B

Content, Tasks, and Response Categories of the Questions

Communication

Question	Content and Task	Major Response Categories
111. Oil Refining	Information given in verbal and numerical form is to be expressed schematically.	1. An accurate and complete diagram is given. 2. A correct numerical answer is given without a diagram. 3. The diagram is inaccurate or incomplete.
112. Verbiants	A set of mathematical relationships described verbally is to be expressed with mathematical symbols. The relationships include subtracting, squaring, dividing, and proportionality.	1. A correct mathematical expression is given. 2. A partially correct mathematical expression is given. 3. The expression is incorrect.
113. Equation Translation	A given equation involving squaring, multiplying, dividing, and proportionality is to be expressed verbally.	1. The verbal statement is correct. 2. The statement includes minor errors. 3. The statement is seriously wrong.
114. Schrodinger	A paragraph written by Erwin Schrodinger describing the statistical nature of the kinetic theory of gases is to be rewritten concisely in language understandable to high school seniors.	1. The student's paragraph is clear, concise, and essentially complete. 2. The response is correct and understandable but padded with extraneous comment. 3. The response is inaccurate, incomplete, or vague. 4. The response simply rearranges Schrodinger's words.

Communication

Question	Content and Task	Major Response Categories
122. Line Selection	The student is to select from two versions of a short poem the one that is superior in both sound and sense, explaining the reasons for the choice.	1. The reasons given consider both sound and sense with specific reference to the relations among the words. 2. The reasons consider either sound or sense effectively but not both. 3. Sound or sense or both are mentioned but without being related to the poem.
124. Paragraph Style	A one-paragraph synopsis of The Cask of Amontillado written in eleven short, choppy sentences is to be rewritten in three or four smooth sentences that retain the important content.	1. The paragraph is concise, accurate, graceful, and essentially complete. 2. The paragraph meets three of the four requirements above. 3. The paragraph meets two of the four requirements. 4. The paragraph meets fewer than two of the requirements.
133. Life on Pluto	The student, who is among the first people to land on Pluto, is given a list of statements about the physical, cultural, and technological environment of Pluto and is to put them into a concise paragraph in ordinary language for transmission to Earth.	1. The paragraph integrates most of the information into a coherent paragraph appropriate for the general public. 2. The paragraph is padded with extraneous material. 3. The paragraph is awkward, choppy, convoluted, or a simple paraphrase of the information. 4. The paragraph is limited in content. 5. The paragraph is florid or pretentious.

Communication

Question	Content and Task	Major Response Categories
142. Historical Thinking	The student is to expand clearly, concisely, and with an additional example on the statement, "Historical thinking has its own historical setting," which is exemplified by the differences in views of post-Civil War historians in South Carolina and in Massachusetts.	1. The response is clear and concise with an apt illustration. 2. The response is accurate but over-elaborate, ponderous, or verbose. 3. The response is incomplete or inaccurate in the statement of the principle or in its illustration. 4. The response contains serious errors of structure or diction.
143. Nature of Government	Two paragraphs written in 1789 about the Constitution of Pennsylvania are to be rewritten in clear, concise, contemporary English	1. The paragraph is largely complete, accurate, clear, and fluent. 2. The paragraph is understandable and complete but stilted or verbose. 3. The response is incomplete, awkward, or vague.
144. Study of History	The student is to describe clearly and concisely the relative merits of studying history chronologically by geographic area or by themes, such as "Women in history".	1. Merits of both approaches are described clearly and smoothly. 2. Both approaches are discussed but the diction is poor or the sentence structure is awkward. 3. The response is incomplete or imprecise. 4. The response rambles or editorializes.

Analysis

Question	Content and Task	Major Response Categories
215. Hybrid Grasses	The student is asked to explain why two species of grass can occur together in nature without producing hybrids but can be caused to produce fertile hybrids in the laboratory.	1. The response mentions the controlled conditions of the laboratory and a natural environment hostile to the hybrid but not to the original species. 2. The response mentions one of the above two points. 3. The response merely repeats the information given. 4. The response makes statements or assumptions contrary to the given information.
222. Socrates and Diotima	An exchange from Plato's Symposium is presented in which Diotima asserts that neither the wise nor the foolish seek wisdom because the wise have it and the foolish are ignorant of it, and Socrates asks who then seeks wisdom. The students are to describe the dilemma involved in Socrates' question and state the essential element of an answer.	1. The response states the dilemma as the knowledge that people seek wisdom despite Diotima's assertion, and its resolution in the fact that the wise and foolish represent extremes between which many seekers of wisdom can be accommodated. 2. The problem is described as one of definition of wise and foolish. 3. The response questions one or both of the premises that either the wise or the foolish don't seek wisdom. 4. The response states the student's own view about wisdom. 5. The response merely restates the question.

Analysis

Question	Content and Task	Major Response Categories
223. O'Connor Story	The opening paragraph of a short story by Flanney O'Connor is given in which a number of apparently mundane details are stated about a mother and her teen-age son. The student is to describe the mother-son relationship.	1. Both sides of a relationship involving mutual dependence and resentment are described. 2. Only one side of the relationship is described. 3. The response draws inferences well beyond anything in the paragraph.
224. J.S. Mill Quotation	A paragraph of Mill's is given in which an analogy is drawn between the evidence that an object is visible being found in the fact that people see it and the evidence that an object is desirable being found in the fact that people desire it. The student is asked how Mill's argument might best be attacked.	1. The response described the flaw in an analogy between the physical senses and a subjective response. 2. The response misinterprets Mill's argument, assuming that seeing something is proof of existence rather than visibility and that desiring something is proof of universal desirability rather than an individual response. 3. The response deals with the physical sense or a subjective response but not both.
231. Chimpomat	The student is to describe the conclusions to be drawn from an experiment in which chimpanzees learned to work for different kinds of coins that would get different amounts of grapes out of a slot machine. Some chimps would spend all their coins as soon as they got them while others would hide their coins and use them when work for them wasn't available.	1. Two or more correct conclusions are stated, e.g., that chimps learn the symbolic value of coins, that they learn to discriminate among symbols of different value, that they differ in the style of their behavior, and that they are capable of planning. 2. Two or more correct conclusions are stated but one or more faulty conclusions are also stated (e.g., chimps learn that work should be as profitable as possible; chimps learned they would go hungry if they didn't save their coins.

Analysis

Question	Content and Task	Major Response Categories
231. Chimpomat		3. The response merely repeats information given.
232. U.S. and Russia	A quotation is given in which the fact that the industrial production of the U.S. is greater than Russia's despite Russia's greater area and population is attributed to the "climate of freedom" in the U.S. The student is asked to state the most crucial assumption underlying the argument and explain why it is crucial.	1. Any one of several acceptable assumptions is stated and the reason for its being crucial is clearly explained. 2. An acceptable assumption is stated but not explained. 3. An assumption is stated that cannot be inferred from the quotation (e.g., Americans are greedy, Russia doesn't have freedom because it's Communist). 4. The response only restates information already given.
234. Capital Punishment	As adviser to a state legislature considering legislation on capital punishment, the student is to consider five different procedures for gathering data concerning the deterrent value of capital punishment and is to pick the most and least informative procedure and explain both choices.	Any of five procedures may be selected as best or worst. Responses are classified according to the quality of both explanations. Good explanations focus on some specific advantage or disadvantage of the procedure. Poor explanations are irrelevant, perhaps arguing the merits of capital punishment rather than addressing the question, or simply assert that one or another method is best or worst. Intermediate explanations have some merit but are vague or over-general, perhaps stating that a particular method is desirable because it provides quantifiable information.

Analysis

Question	Content and Task	Major Response Categories
244. Oppenheimer	The student is given two paragraphs from the report of the Personnel Security Board that considered charges after World War II that physicist J. Robert Oppenheimer was a security risk. One summarizes the majority view that government security takes precedence over individual rights and that the government can protect the rights of citizens while protecting its own security. The other states the minority view that the government must accept some level of risk to its security in order to protect the rights of citizens. The student is to state the issue illustrated by the two opposing views.	1. The issue is clearly stated to be the continuing struggle between personal liberty and national security. 2. Either national security or personal freedom is discussed, but the two are not opposed. 3. The issue is described as one of defining concepts or setting standards of personal liberty. 4. The response only restates points made in one or both of the paragraphs without identification of the issue.
245. Roman Conquest	The student is given a paragraph describing the Roman conquest of the interior of the Spanish Peninsula and is asked to identify the statements that can be accepted as factual and those that a skeptical historian might justifiably question and the reasons for both the acceptances and the challenges.	1. Accepted statemets refer to things that are observable or verifiable, such as mountainous terrain; challenged statements refer to interpretive or speculative statement, e.g., concerning reasons for different attitudes among different segments of the conquered people. 2. The response is mixed, having both justifiable and questionable selections of statements and reasons for their selection. 3. The response is overly skeptical, challenging almost every statement. 4. The response draws from information not given or adds gratuitous speculation.

Synthesis

Question	Content and Task	Major Response Categories
315. Phytoplankton and dissolved Nitrates	Two curves are given showing the variation during the course of a year in the concentrations of small marine plants and dissolved nitrates in sea water. During spring the phytoplankton (plant) concentration rises and the nitrate concentration falls. During the summer the phytoplankton concentration drops and stays low until the following spring. The nitrate concentration rises during the fall and stays high until the following spring. The student is asked to speculate about what kind of process might produce the relationship between the two curves.	1. The response notes the seasonal changes, the cyclic nature of both processes, and a probable connection between them, such as the increased phytoplankton concentration depleting the nitrate supply. 2. The response correctly describes the fluctuations of the two curves but does not relate them. 3. The response includes errors in reading the curves. 4. The response asserts that a relationship does not exist or cannot be inferred.
321. Creative Endeavor	A quotation from Schoenberg is given on the importance of several elements of form or organization in music. The student is to describe how those elements might apply in short stories as well as music.	1. Schoenberg's elements of form--development, growth, elaboration, fate--are described as they might apply to short stories as well as music. 2. The response discusses organization but does not draw its content from Schoenberg. 3. The response merely restates Schoenberg's elements and asserts their importance for short stories.
322. Litton Industries	Litton Industries is described as a major industrial corporation, and a paragraph is quoted from its annual report drawing an analogy between Litton Industries and stained glass--both merge technical and artistic skill and form	1. At least one additional quality is given that could be common to Litton and stained glass. 2. The same qualities given are elaborated.

Synthesis

Question	Content and Task	Major Response Categories
322. Litton Industries	a unified structure from a diversity of shapes and colors. The student is to extend the analogy, suggesting another possible parallel between Litton and stained glass.	3. A quality is described that characterizes either Litton or stained glass but not both. 4. Inappropriate images or analogies are given.
325. Yin-yang	The Taoist symbol is given and the student is asked to describe one idea the symbol conveys and what elements of the symbol produce that impression.	1. A general concept--e.g., good and evil, balance, harmony, unity, change--is stated and elaborated, and its origin in the shapes and colors is described. 2. The concept is simplistic, unelaborated. 3. The source of the concept in elements of the symbol is not described. 4. A specific association rather than a general concept is given.
332. Suicide and Marriage	A bar graph is given showing suicide rates in the U.S. for single, married, widowed and divorced men and women in two age groups. The student is to draw inferences from the graph about social conditions for men and women in the U.S. and support them through references to the graph.	1. A defensible inference is given with appropriate reference to information in the graph. Or, the response points out legitimate problems in drawing inferences about social conditions from suicide rates. 2. One or more defensible inferences are given without explanation. 3. The response correctly restates information given but does not reach any inferences. 4. The response is based on an erroneous reading of the graph.

Synthesis

Question	Content and Task	Major Response Categories
332. Suicide and Marriage		5. The response over-interprets the data, drawing unjustified inferences, or adds information not given in the graph.
342. Governmental Strengths and Weaknesses	Fifteen characteristics of the federal government are listed that refer to the three branches of government and the elaboration of checks and balances. The student is to describe both a strength and a weakness of the federal government suggested by the list.	1. The response describes the value of checks and balances and the problems of divided responsibility and authority that accompany them. 2. The answer makes a valid statement about the federal government, but does not relate it clearly to the given statements. 3. A single statement in the list is discussed without a general statement being made.

Awareness

412. Earth's Rotation	The students are asked to consider a proposal for producing energy that would result in a slowing of the earth's rotation and to suggest several questions that should be raised in the inevatable Environmental Impact Report. The questions raised need not be answered.	Classification of responses is somewhat elaborate. It is based on the relative number of good, acceptable, and poor questions, which keeps the sheer number of questions raised from affecting the scoring of the responses. Good questions are specific enough to reflect knowledge of a probable environmental effect, e.g., a change in climate due to changes in atmospheric heat flow resulting from longer days and nights. Acceptable questions raise a general issue, e.g., a change in climate, that is unelaborated or too vague to indicate the level of the student's awareness. Poor questions are

Awareness

Question	Content and Task	Major Response Categories
412. Earth's Rotation		trivial, overly broad, or unrelated to a change in the earth's rotation, e.g., who will decide whether to try the proposed scheme.
413. Ephemeral Insects	A passage is quoted about a race of insects who live only an hour and therefore experience only a portion of the sun's daily movement. As evening approaches, the insects' wisest scientist explains that science has established through generations of experience that the sun moves only westward and that it is therefore soon to disappear forever. The student is asked to describe what the passage illustrates about the nature of science.	1. The response mentions the dependence of science on experience and the need for revision to accommodate new kinds of experience or observations. 2. The response mentions only the need for revision of scientific understanding from one generation to the next, without mention of its dependence on experience. 3. The response discussed science in terms unrelated to the passage.
424. Haiku	The student is to describe the meaning of the following poem. I must roll over. Beware of local earthquakes Bedfellow cricket.	1. The response mentions the interconnectedness of life and concern for the effects of one's actions on others. 2. The response mentions one but not both of the above points. 3. The response is very broad or vague but pertinent, an allusion to nature or life in a broad sense. 4. The interpretation is quite literal, e.g., referring to animals' foreknowledge of earthquakes.

Awareness

Question	Content and Task	Major Response Categories
431. Herbert Quote	The student, commissioned to produce a documentary film on the lives of the elderly poor in the inner city for showing to high school audiences, is to explain probable difficulties to be expected in view of the following quotation of George Herbert. "One half the world knows not how the other half lives."	1. The response describes the difficulty in understanding experiences markedly different from one's own and cites specific areas of difference between elderly poor and most high school students. 2. The response makes the general point, without elaboration or illustration, that differences in experience make understanding difficult. 3. The response denies that any problem should exist.
436. Canada	The student is to give probable reasons why a strike by pilots and air-traffic controllers in bilingual Quebec was successful in preventing a shift from English to French as the language for use between pilots and controllers.	1. The response includes references to the need for rapid, effective communication, its importance for human safety, and the exercise of collective power by a unified group. 2. The response is incomplete with respect to one or more of the above points. 3. The response editorializes, brings in extraneous arguments, or is not pertinent.
441. Presidential Candidate	The following statement by a representative of a group of scientists and engineers supporting a particular presidential candidate is given. "By the time we were through, any guy in Pittsburgh in a T-shirt with a can of beer in his hand knew the smartest people in this country considered _____ to be unfit (for the presidency)."	1. The response states that an elitist political system would develop. 2. The response describes broad, general changes only loosely if at all related to elitism. 3. The response denies that any change would occur. 4. The assumptions are discussed rather than the changes that might occur.

Awareness

Question	Content and Task	Major Response Categories
441. Presidential Candidate	The student is to describe changes that might occur in our political institutions or processes if the assumptions underlying the above quotation were widely accepted.	
444. Poor and Mean	The student is to draw some inferences about the society in which the following verse was written. "Though I am but poor and mean, I will move the rich to love me, If I'm modest, neat and clean, And submit when they reprove me."	1. The response makes an explicit statement about one or more values implied in the poem--propriety, submissiveness, material wealth--going beyond the specific words of the poem, and at least implies the existence of a strong class structure. 2. The response points out the class distinctions but makes no reference to values. 3. The response focuses on the writer's individual position without mention of the society in which he writes.

Notes:

1. The first digit of the three-digit question number indicates the competence being assessed according to the following code.

    1. Communication
    2. Analysis
    3. Synthesis
    4. Awareness

    The second digit indicates the content field according to the following code.

    1. Natural Sciences
    2. Humanities
    3. Social Sciences
    4. History and Political Science

2. Most of the questions have six response categories, although a few have five and a few have up to nine. Only the primary distinctions in the response categories are described here. The others are usually varying degrees of the categories described, for example, involving a greater number or more serious errors.